Influence des points d'ancrage pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés

Emmanuel Ep Prochasson, Emmanuel Morin

Influence des points d'ancrage pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés

Emmanuel Prochasson Emmanuel Morin
Université de Nantes, LINA - UMR CNRS 6241
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
{emmanuel.prochasson,emmanuel.morin}@univ-nantes.fr

Résumé. L'extraction de lexiques bilingues à partir de corpus comparables affiche de bonnes performances pour des corpus volumineux mais chute fortement pour des corpus d'une taille plus modeste. Pour pallier cette faiblesse, nous proposons une nouvelle contribution au processus d'alignement lexical à partir de corpus comparables spécialisés qui vise à renforcer la significativité des contextes lexicaux en s'appuyant sur le vocabulaire spécialisé du domaine étudié. Les expériences que nous avons réalisées en ce sens montrent qu'une meilleure prise en compte du vocabulaire spécialisé permet d'améliorer la qualité des lexiques extraits.

Abstract. Bilingual lexicon extraction from comparable corpora gives good results for large corpora but drops significantly for small size corpora. In order to compensate this weakness, we suggest a new contribution dedicated to the lexical alignment from specialized comparable corpora that strengthens the representativeness of the lexical contexts based on domain-specific vocabulary. The experiments carried out in this way show that taking better account the specialized vocabulary induces a significant improvement in the quality of extracted lexicons.

Mots-clés : Corpus comparable, extraction de lexiques bilingues, points d'ancrage.

Keywords: Comparable corpus, bilingual lexicon extraction, anchor points.

1 Introduction

L'extraction de lexiques bilingues à partir de corpus comparables a connu un essor important depuis le début des années quatre-vingt dix (Rapp, 1995; Fung, 1998; Rapp, 1999; Chiao & Zweigenbaum, 2002; Morin & Daille, 2004, entre autres). Cet intérêt pour l'exploitation de corpus comparables est principalement lié aux difficultés de disposer de corpus parallèles, notamment lorsqu'il s'agit d'exploiter un matériau textuel ne faisant pas intervenir l'anglais. En outre, les lexiques bilingues obtenus à partir de corpus parallèles sont quelque peu biaisés. En effet, un corpus parallèle étant constitué d'un texte dans une langue source et de sa traduction dans une langue cible, le vocabulaire rencontré dans la partie traduite est fortement influencé par celui de la langue source en particulier dans les domaines spécialisés. Les recherches se sont donc tournées vers l'exploitation de corpus comparables, c'est-à-dire des corpus regroupant des textes dans des langues différentes qui ne sont pas en correspondance de traduction mais qui partagent des traits communs comme le domaine, le thème, la période... Ces corpus sont plus largement disponibles, et les textes qui les composent ont été écrits indépendamment dans chaque langue.

La méthode par traduction directe associée à l'exploitation des corpus comparables donne de bons résultats pour des corpus volumineux (Rapp, 1999, centaine de millions de mots) mais chute fortement pour des corpus d'une taille plus modeste (Chiao & Zweigenbaum, 2002, autour de 1 million de mots). Nous proposons dans cet article une nouvelle contribution à la méthode par traduction directe visant à améliorer la qualité des lexiques bilingues extraits à partir de corpus comparables spécialisés de taille modeste. Cette contribution vise à renforcer la significativité des contextes lexicaux de la méthode par traduction directe en s'appuyant sur le vocabulaire spécialisé caractéristique du domaine étudié.

Après avoir présenté la méthode par traduction directe et précisé les enjeux de la caractérisation des contextes des termes en section 2, nous précisons la notion de point d'ancrage sous-jacente à ce travail et montrons son intégration avec la méthode directe en section 3. Cette notion est illustrée à travers deux exemples de vocabulaire spécialisé : les translittérations et les composés savants. Nous présentons ensuite les ressources et outils utilisés dans ce travail en section 4, puis les différentes expériences réalisées visant à préciser l'influence de ces points d'ancrage dans le processus d'alignement en section 5. Enfin, la section 6 dresse le bilan de ce travail.

2 Extraction lexicale bilingue à partir de corpus comparables

2.1 Caractérisation des contextes des termes

Les premières recherches en extraction lexicale à partir de corpus comparables se sont naturellement éloignées des approches proposées pour les corpus parallèles et ont proposé de chercher des caractéristiques qui seront proches entre un terme i et sa traduction $t(i)$ et éloignés entre un terme i et d'autres mots qui n'en sont pas traductions. Fung (1995) a d'abord proposé de s'intéresser à la *productivité* des termes, c'est-à-dire le nombre de voisins directs rencontrés dans le corpus ; l'hypothèse posée étant que ces productivités (productivités à gauche et à droite) étaient proches entre un terme i et sa traduction $t(i)$, mais très différentes pour des termes qui ne sont pas en rapport de traduction. À la même période, Rapp (1995) propose une approche comparant les motifs de relations d'associations entre un terme i et l'ensemble de ses voisins (l'association entre deux termes indique dans quelle mesure ils apparaissent plus souvent ensemble que *par chance*). Il proposait l'hypothèse qu'un mot i et sa traduction $t(i)$ devaient avoir des motifs d'associations comparables.

Ces deux approches réalisent le bond conceptuel vers l'extraction à partir de corpus comparables : elles proposent de s'appuyer non plus sur un terme i à traduire, mais sur son *contexte*, c'est-à-dire sur l'ensemble des mots avec lesquels i cooccurre. Elles font ainsi écho à la proposition de Firth (1957) : « *on reconnaît un mot à ses fréquentations* » [1]. Le processus d'extraction à partir de corpus comparables repose donc en grande partie sur la capacité à caractériser les contextes des termes à aligner. La méthode par traduction directe propose de s'appuyer sur des ressources linguistiques existantes pour comparer les contextes des termes. Elle propose aussi une structure de données pour enregistrer ces contextes.

2.2 Méthode par traduction directe

La méthode par traduction directe, introduite par Fung (1998) et Rapp (1999), s'inspire des modèles de contextes pour les corpus monolingues ainsi que des méthodes d'extraction d'information en remplaçant les concepts de *requête* et de *document* par ceux de *termes* et *contexte*

[1] « *You shall know a word by the company it keeps* ».

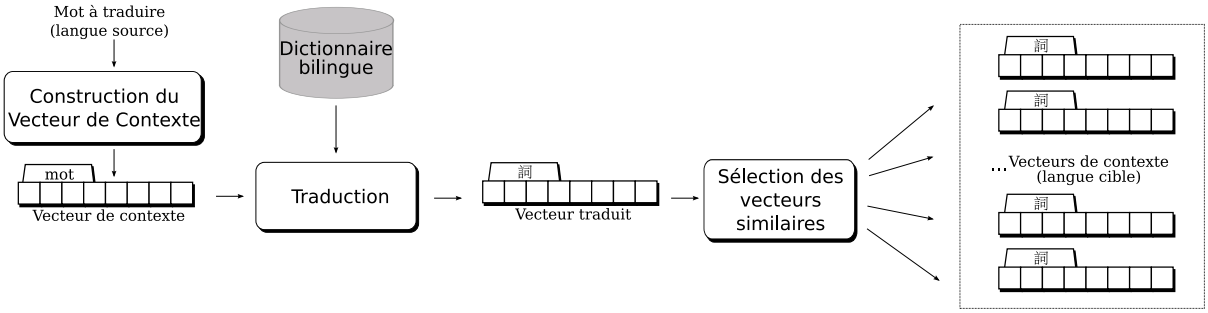

FIG. 1 – Méthode par traduction directe.

des termes. Le principe de cette méthode, illustrée par le schéma de la figure 1, s'implémente de la manière suivante :

Construction des vecteurs de contexte pour chaque mot i. Nous collectons tous les mots qui cooccurrent dans une fenêtre donnée. Nous obtenons, pour chaque mot i des corpus source et cible, un *vecteur de contexte* qui regroupe l'ensemble des mots j qui apparaissent avec i, associés avec leur nombre de cooccurrences. Nous appelons i la *tête* du vecteur et j les *éléments* du vecteur. Et nous désignons par $occ(i, j)$ le nombre de cooccurrences des éléments i et j.

Normalisation des vecteurs de contexte en utilisant des *mesures d'association*, telles que le Taux de vraisemblance (Dunning, 1993) (cf. équation 1 et table 1), entre la tête d'un vecteur et ses éléments. De cette manière, les vecteurs de contexte sont comparés non plus sur les fréquences de leurs éléments, mais sur le motif des associations entre la tête et les éléments des vecteurs.

Traduction des vecteurs de contexte en utilisant un dictionnaire bilingue. Pour chaque mot dont nous voulons obtenir la traduction, nous traduisons les éléments de son vecteur de contexte. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de i l'ensemble des traductions proposées (lesquelles sont pondérées par la fréquence de la traduction en langue cible).

Sélection des vecteurs de contexte proches en utilisant des mesures de similarité. Plus deux vecteurs de contexte sont proches, plus il est probable qu'ils correspondent à des traductions.

Nous obtenons ainsi, pour chaque mot à traduire, une liste ordonnée (par ordre de similarité) des candidats à la traduction.

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

TAB. 1 – Table de contingence pour un couple i et j

$$\begin{aligned}\lambda(i,j) \quad = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) + \\ & (a+b+c+d)\log(a+b+c+d) - (a+b)\log(a+b) - \\ & (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d)\end{aligned} \tag{1}$$

2.3 Quelques résultats de la méthode directe

Il est malheureusement assez difficile de comparer les résultats entre les différentes études publiées sur ce sujet, en raison des différences entre les corpus exploités (en particulier leurs contraintes de construction et leur volume) mais aussi de la couverture et de la pertinence des ressources linguistiques utilisées pour la traduction. À ce jour, il n'existe à notre connaissance aucune expérience et aucun jeu de ressources de référence.

Les résultats de la méthode par traduction directe s'évaluent sur le nombre de candidats correctement alignés, trouvés dans la liste des x premiers candidats proposés (le Top_x). Rapp (1999) obtient par exemple 72 % de résultats corrects pour le Top_1 et 89 % pour le Top_{10} avec un corpus comparable composé d'articles de journaux (135 millions de mots pour la partie anglaise et 163 millions pour la partie allemande) et un dictionnaire bilingue contenant 16 380 entrées (termes simples). Chiao & Zweigenbaum (2002), en s'appuyant sur un corpus médical français/anglais de 1,2 millions de mots et un dictionnaire spécialisé de 18 437 entrées, obtiennent 20 % de précision pour le Top_1 et environ 60 % pour le Top_{20}. Ces résultats sont moins bons que ceux de Rapp (1999) mais s'expliquent aisément par la différence de taille des corpus utilisés. Précisons que nous nous intéressons dans cet article uniquement à l'alignement de termes simples mais que d'autres recherches se sont portées sur l'alignement de termes complexes, notamment Morin & Daille (2004).

3 Points d'ancrage dans les corpus comparables

3.1 Vocabulaire spécialisé comme point d'ancrage

Nous cherchons à renforcer la significativité des vecteurs de contexte construits sur des petits corpus, en recherchant des *points d'ancrage*. Ces points d'ancrage doivent être des *éléments de confiance*, c'est-à-dire des éléments dont l'absence ou la présence dans un vecteur de contexte est particulièrement discriminante pour caractériser un terme. Notons que cette notion est proche de celle utilisée avec les corpus parallèles (Brown *et al.*, 1991), c'est-à-dire des éléments alignés avec confiance et sur lesquels les méthodes peuvent s'appuyer pour aligner leurs voisins.

En pratique, ces points d'ancrage doivent avoir plusieurs propriétés :

1. Ils doivent être faciles à identifier.
2. Ils doivent être pertinents, relativement aux thèmes des documents à aligner.
3. Ils doivent être peu polysémiques, pour ne pas être ambigus.

Nous émettons l'hypothèse que ces points d'ancrage sont des éléments discriminants dans la caractérisation des contextes des termes et qu'ils peuvent être employés efficacement pour améliorer les résultats de la méthode directe. Dans ce cadre, nous nous sommes intéressés aux *translittérations* et aux *composés savants*.

Nous appelons *translittération* le phénomène d'emprunt d'un mot d'une langue source à destination d'une langue cible, ne partageant pas nécessairement les mêmes phonèmes ni les mêmes symboles d'écriture. Le mot emprunté est alors adapté graphiquement dans la langue cible, sur la base de sa prononciation et non de son sens (Knight & Graehl, 1997). Dans Prochasson *et al.* (2008), nous avons montré la proéminence des translittérations dans le corpus anglais-français-japonais que nous utilisons ici et que nous présentons en section 4.1. Elles sont de

bonnes candidates pour être des points d'ancrage. Elles sont en effet faciles à identifier, car écrites dans un syllabaire principalement dédié aux mots d'emprunt en japonais (les *katakanas*) ; elles sont également représentatives d'un vocabulaire spécifique qui recouvre le vocabulaire spécialisé (elles sont parfois utilisées alors même qu'il existe un mot plus ancien et équivalent en japonais). Les translittérations japonaises sont issues pour la plupart de l'anglais, mais peuvent dans de nombreux cas être alignées avec des termes français, en raison des relations de *cognats* fréquentes entre le français et l'anglais. Par exemple, le terme japonais インスリン / `i-n-su-ri-n` peut s'aligner en anglais avec *insulin* et en français avec *insuline*.

Nous nous sommes également penchés sur les *composés savants*. Ils s'agit de mots, en français et en anglais, construits à partir de racines spécifiques (Namer, 2005). Claveau (2007), s'intéressant à la traduction automatique de termes biomédicaux observe que « *les termes biomédicaux sont construits sur les mêmes racines grecques et latines, et leurs dérivations très régulières* » (p. 2). Ces mots paraissent être des points d'ancrage pertinents dans le cas d'un corpus spécialisé sur le *diabète et l'alimentation* tel que le nôtre, puisqu'ils sont représentatifs d'un vocabulaire *savant* et qu'ils peuvent être facilement identifiés à partir de leur morphologie.

3.2 Modification de la méthode directe

Nous avons choisi de modifier la méthode par traduction directe en accordant plus d'importance aux points d'ancrage lors du calcul de l'association entre la tête d'un vecteur et ses éléments. Après avoir calculé l'association de façon standard, nous rehaussons le score des points d'ancrage et diminuons le score des autres éléments de manière à ce que les sommes des scores initiaux et finaux soient identiques (voir équations 2 à 4).

$$assoc_a^v := assoc_a^v + \beta \ , \ si\ a \in PA \tag{2}$$

$$assoc_a^v := assoc_a^v - Decalage \ , \ si\ a \notin PA \tag{3}$$

$$Decalage := \frac{\#PA}{\#\neg PA} \times \beta \tag{4}$$

Dans ces équations, $\#PA$ est le cardinal de l'ensemble des points d'ancrage extraits (et $\#\neg PA$ le cardinal des autres éléments) et $assoc_a^v$ est la mesure d'association de l'élément a dans le vecteur de contexte v. Le paramètre β permet de calibrer l'importance donnée aux points d'ancrage. Ce paramètre est ajouté de façon absolue au score de chaque point d'ancrage, et non de façon proportionnelle par rapport au score initial. En effet, nous souhaitons rehausser le score de tous les points d'ancrage de manière à ce qu'ils soient pris en compte en priorité par les mesures de similarité.

4 Ressources et outils exploités

4.1 Ressources linguistiques

Dans le cadre de cette étude, nous avons constitué un corpus trilingue, français, anglais et japonais. Les documents qui composent ce corpus ont été extraits du Web et concernent le thème *alimentation et diabète* et appartiennent au registre *scientifique*[2]. Les documents ont été sélec-

[2]Documents écrits par des experts à destination d'autres experts (Pearson, 1998, p. 36).

tionnés manuellement, à partir de requêtes précises sur des moteurs de recherche, mais aussi en suivant les liens internes des pages proposées. Ces documents ont été convertis de leur format source HTML ou PDF en texte brut, puis normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique et lemmatisation. Nous avons ainsi collecté 257 000 mots pour la partie française, 235 000 pour la partie japonaise et 250 000 mots pour la partie anglaise.

Le dictionnaire français-japonais, nécessaire pour l'étape de traduction de la méthode directe, est composé de quatre dictionnaires disponibles librement sur Internet[3] ainsi que du *Dictionnaire scientifique français-japonais* (1989). Il contient 173 156 entrées, dont 114 461 sont des termes simples, avec une moyenne de 2,1 traductions par entrée. Nous avons utilisé le dictionnaire *JMDict* pour l'anglais/japonais[4] qui est disponible librement sous une licence *Creative-Commons* (Attribution-ShareAlike). Nous l'avons complété de traductions de termes techniques issus de différentes sources : dictionnaire du Ministère de l'Éducation japonais et du *National Institute of Informatics* (Tokyo)[5] ainsi que du *Dictionary of Technical Term* (Kotani & Kori, 1990). Il contient 589 956 entrées avec une moyenne de 2,3 traductions par entrée et seulement 49 208 termes simples.

Pour évaluer la qualité de l'extraction, nous avons construit une liste de traductions connues. Nous avons extrait du corpus les mots français et anglais les plus fréquents ($N_{occ} > 50$) dont la traduction en japonais est connue. Parmi ces traductions, nous avons sélectionné celles apparaissant fréquemment dans le corpus japonais ($N_{occ} > 50$) pour construire une liste de 98 traductions français-japonais et 99 traductions anglais-japonais. Ce protocole pour constituer une liste de termes utilisée pour l'évaluation est semblable à celui présenté dans Chiao & Zweigenbaum (2002) qui travaillent avec une liste de 95 termes pour un corpus anglais/français de 1,2 millions de mots.

4.2 Outils

De nombreux efforts ont été réalisés pour aligner automatiquement les translittérations. Dans ce travail, nous avons utilisé un outil réalisant la détection automatique de translittérations entre l'anglais et le japonais (Tsuji, 2001). Bien que plus simple que l'algorithme proposé par Knight & Graehl (1997), cet outil, basé sur des chaînes de Markov, donne des résultats satisfaisants. Il génère un ensemble de correspondances potentielles pour une entrée donnée en katakana ou en anglais. Les résultats doivent alors être comparés avec un vocabulaire existant pour sélectionner les candidats les plus probables. Nous obtenons avec cet outil 589 paires de translittérations pour le couple anglais/japonais. En ce qui concerne la détection des translittérations entre le français et le japonais, nous avons utilisé le même outil, n'ayant pu obtenir un outil efficace dédié. Avant traitement, les termes français à comparer sont normalisés pour faire disparaître les signes diacritiques spécifiques (mais ils sont réintégrés dans les couples alignés). Nous obtenons 526 paires de translittérations pour le couple français/japonais.

En ce qui concerne la détection des composés savants, nous nous sommes appuyés sur une liste de 606 affixes médicaux utilisés en anglais avec leur correspondance en japonais[6]. Le processus

[3] kanji.free.fr ; quebec-japon.com/lexique/index.php?a=index&d=25 ; dico.fj.free.fr/index.php ; quebec-japon.com/lexique/index.php?a=index&d=3

[4] www.csse.monash.edu.au/~jwb/j_jmdict.html

[5] sciterm.nii.ac.jp/cgi-bin/reference.cgi

[6] www.medo.jp/a.htm

d'extraction est trivial : en compilant une expression régulière par affixe, il cherche les mots anglais correspondant dans les dictionnaires bilingues utilisés pour l'alignement. Les mots extraits sont conservés ainsi que leurs traductions en japonais, pour obtenir des paires de traductions utilisées comme points d'ancrage dans les vecteurs de contexte. La liste des affixes correspond aux racines anglaises, mais elle peut facilement être traduite en français, en accord avec la remarque de Claveau (2007). Nous nous sommes inspirés de ce travail pour écrire quelques règles simples de conversion. La terminaison `-y` (comme dans *psychology*) est par exemple transformée en `-ie` en français (*psychologie*). Certains affixes sont très productifs (typiquement le privatif `a-`) et les mots correspondants extraits ne sont pas forcément formés à partir de ces préfixes (le `a` de *armoire* ne correspond pas au privatif). Tous les affixes générant plus de 1 000 correspondances sur les ressources ont donc été écartés. Ils sont toutefois assez rares et 12 seulement ont été écartés pour l'anglais, 17 pour le français. Nous avons ainsi obtenu 17 210 composés savants en anglais, correspondant à 60 341 traductions (les ressources linguistiques comprennent des traductions multiples pour un seul élément source). Nous avons également obtenu 8 254 composés savants français, soit 24 240 traductions. Ces différences de résultats proviennent de la nature des dictionnaires bilingues.

5 Expériences et résultats

5.1 Protocoles

Nous avons réalisé plusieurs expériences pour évaluer l'efficacité et l'impact des points d'ancrage pour l'alignement lexical bilingue :
– (*a*) : approche standard utilisant simplement la méthode par traduction directe ;
– (*b*) : en utilisant les translittérations détectées automatiquement ;
– (*c*) : en utilisant les mots savants extraits automatiquement.
L'expérience *a* est une expérience témoin, utilisée comme étalon pour être comparée avec les expériences *b* et *c*.

Les expériences sont réalisées avec les mêmes paramètres. Nous utilisons le *Taux de vraisemblance* comme mesure d'association et le *Cosinus* comme mesure de similarité. Nous utilisons une fenêtre de taille 25 mots avant et après la tête pour la constitution des vecteurs de contexte, et un nombre d'occurrences minimal de 3 pour qu'un mot soit pris en compte. Ces paramètres sont ceux qui donnent les meilleurs résultats pour l'expérience témoin. Dans le cas des expériences *b* et *c*, nous faisons varier le paramètre β entre 1 et 10 (par pas de 1).

5.2 Résultats

La table 2 synthétise les résultats obtenus pour les expériences *a*, *b* et *c* pour les Top 1 et 10, pour l'alignement anglais-japonais et français-japonais (entre crochets, le gain obtenu).
Les résultats pour l'expérience de contrôle *a* sont comparables à ceux obtenus par Chiao & Zweigenbaum (2002) discutés en section 2.3. Dans le cas de l'anglais, le gain obtenu en s'appuyant sur les points d'ancrage est important : à hauteur de 12,5 % en utilisant les translittérations (exp. $b - Top_1$) et 18,8 % en utilisant les mots savants (exp. $c - Top_1$). Le gain est moins important pour le français-japonais : il est nul pour le Top_1 en utilisant les translittérations et atteint 10 % en utilisant les composés savants. La moins bonne qualité de résultats avec le fran-

	a	*b*	*c*
Anglais/Japonais (Top_1)	17,1 %	20,2 % [18,2 %]	20,2 % [18,2 %]
Anglais/Japonais (Top_{10})	36,3 %	39,3 % [8,2 %]	40,4 % [11,2 %]
Français/Japonais (Top_1)	20,4 %	20,4 % [0,0 %]	22,4 % [10,0 %]
Français/Japonais (Top_{10})	36,7 %	37,8 % [2,8 %]	38,8 % [5,6 %]

TAB. 2 – Résultats de l'alignement anglais-japonais et français-japonais ($\beta = 8$)

çais peut s'expliquer par la moins bonne qualité des listes de points d'ancrage. En particulier, les translittérations ont été extraites avec un outil initialement dédié à la détection entre l'anglais et le japonais, sans oublier que les translittérations entre le français et le japonais sont plus rares.

5.3 Influence des points d'ancrage

La figure 2 compare les résultats obtenus entre l'expérience témoin et l'utilisation des composés savants, dans le cas de l'alignement français-japonais ($\beta = 8$). Elle présente l'évolution des positions des traductions correctes dans les listes de candidats obtenues à la fin du processus d'alignement (en ordonnée), ainsi que de leurs scores de similarité (en abscisse).

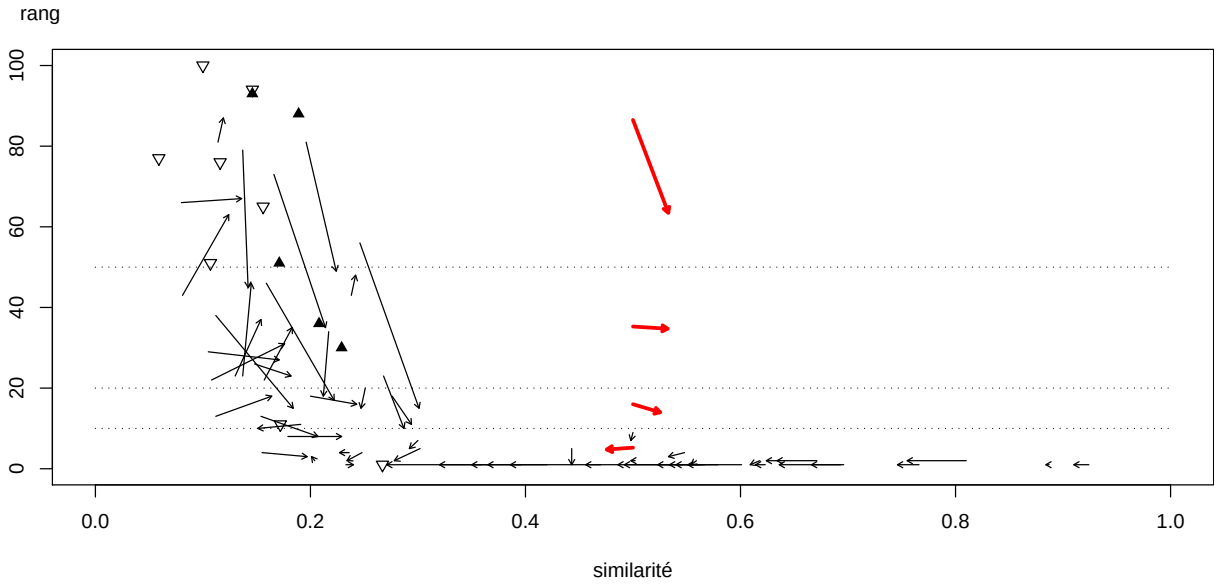

FIG. 2 – Rang et score des traductions correctes pour l'alignement français-japonais ($\beta = 8$)

Les triangles vides représentent les traductions n'étant plus obtenues avec l'utilisation des composés savants, alors que les triangles noirs indiquent les nouvelles traductions obtenues, indisponibles dans le cas de l'expérience témoin. Les flèches fines représentent le déplacement d'une traduction entre l'expérience témoin (début de la flèche) et l'expérience utilisant des points d'ancrage (pointe de la flèche). Enfin, les quatre flèches plus épaisses sont la somme des flèches fines pour chaque zone délimitée par des pointillés.

Cette figure montre d'abord que le nombre de traductions introduites est proche du nombre de traductions disparues. Elles correspondent à des traductions instables, étant très sensibles aux différents paramètres utilisés (taille de fenêtre, mesure d'association et de similarité...). Les flèches permettent de mieux comprendre l'influence des points d'ancrage. Elles indiquent en effet que, en moyenne, les traductions correctes obtiennent un meilleur rang dans les résultats

de l'alignement. C'est particulièrement visible pour les traductions initialement mal classées (Top_{50} à Top_{100}). Leur rang est largement amélioré comme l'indique la somme des vecteurs pour cette zone. Ce constat est valable pour les autres zones, même s'il est moins flagrant. Dans tous les cas, en moyenne, l'utilisation des points d'ancrage améliore le classement des traductions correctes dans la liste des candidats obtenus. Toutefois, les traductions initialement correctement alignées (Top_{10} ou inférieur) sont peu reclassées (elles ne sont toutefois pas désavantagées, même si leur indice de similarité moyen diminue). Ces observations viennent compléter les résultats présentés : elle montre une tendance au réarrangement des candidats à la traduction vers des positions plus avantageuses, quelque soit leur rang initial.

6 Conclusion

Nous avons proposé dans cet article une nouvelle contribution à l'extraction lexicale bilingue à partir de corpus comparables. Notre hypothèse repose sur l'identification et la confiance en un vocabulaire spécialisé pour améliorer les résultats de la méthode directe. Cette hypothèse a été confirmée par l'expérience : nous avons montré que les résultats sont améliorés pour les Top_1 et Top_{10}, mais aussi que le classement des candidats à la traduction est notablement et globalement amélioré en utilisant des points d'ancrage.

Cette étude ouvre la voie à de nouvelles perspectives pour améliorer les méthodes d'alignement lexical à partir de corpus comparables spécialisés d'une taille réduite. D'un côté, la qualité des points d'ancrage que nous avons extraits peut être améliorée en utilisant des outils plus performants. De l'autre, il est probable qu'il existe d'autres points d'ancrage pertinents, dont l'extraction est permise par des travaux transversaux en traitement automatique des langues naturelles (typiquement les cognats).

Enfin, cette étude invite à repenser la façon dont sont caractérisés et comparés les contextes des termes. En effet, la méthode par traduction directe s'appuie sur la comparaison des associations entre les têtes et les éléments des vecteurs de contexte, d'une langue vers une autre. Nous l'avons modifiée en utilisant un score artificiel, n'étant plus une mesure d'association et avons obtenu des résultats prometteurs. En parallèle à la recherche et à l'exploitation de nouveaux points d'ancrage, nous souhaitons donc réfléchir à de nouvelles mesures discriminantes, plus adaptées que ne le sont les mesures d'association pour l'extraction lexicale bilingue à partir de corpus comparables spécialisés d'une taille réduite.

Remerciements

Les auteurs tiennent à remercier Kyo Kageura (Université de Tokyo) et Akiko Aizawa (*National Institute of Informatics*, Tokyo) pour leur aide et leurs conseils avisés. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009.

Références

BROWN P. F., LAI J. C. & MERCER R. L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics (ACL'91)*, p. 169–176, Berkeley, CA, États-Unis d'Amérique.

CHIAO Y.-C. & ZWEIGENBAUM P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, p. 1208–1212, Tapei, Taiwan.

CLAVEAU V. (2007). Inférence de règles de réécriture pour la traduction de termes biomédicaux. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'07)*, p. 111–120, Toulouse, France.

DUNNING T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1), 61–74.

FIRTH J. (1957). *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

FUNG P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In D. YAROVSKY & K. CHURCH, Eds., *Proceedings of the 3rd Workshop on Very Large Corpora (VLC'95)*, p. 173–183, Somerset, NJ, États-Unis d'Amérique.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction : From Parallel Corpora to Non-parallel Corpora. In D. FARWELL, L. GERBER & E. HOVY, Eds., *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, p. 1–16, Langhorne, PA, États-Unis d'Amérique.

KNIGHT K. & GRAEHL J. (1997). Machine transliteration. In P. R. COHEN & W. WAHLSTER, Eds., *Proceedings of the 3rd Annual Meeting of the Association for Computational Linguistics (ACL'97) and 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL'97)*, p. 128–135, Madrid, Espagne.

KOTANI T. & KORI A. (1990). *Dictionary of Technical Terms*. Kenkyusha.

MORIN E. & DAILLE B. (2004). Extraction terminologique bilingue à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, **45**(3), 103–122.

NAMER F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, p. 63–72, Dourdan, France.

PEARSON J. (1998). *Terms in Context*. John Benjamins publishing company.

PROCHASSON E., KAGEURA K., MORIN E. & AIZAWA A. (2008). Looking for transliterations in a trilingual english, french and japanese specialised comparable corpus. In *Proceedings of the 1st Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference (LREC'08)*, p. 83–86, Marrakech, Maroc.

RAPP R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, p. 320–322, Cambridge, MA, États-Unis d'Amérique.

RAPP R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 519–526, College Park, MD, États-Unis d'Amérique.

TSUJI K. (2001). Automatic extraction of translational japanese-katakana and english word pairs from bilingual corpora. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Langages (ICCPOL'01)*, p. 245–250, Taichung, Taiwan.